DOCUMENT RESUME

ED 223 429 SE 039 588

AUTHOR O'Brien, Francis J., Jr.
TITLE A Derivation of the Sample Multiple Correlation Formula for Standard Scores.
PUB DATE Nov 82
NOTE 45p.
PUB TYPE Guides - Classroom Use - Materials (For Learner) (051)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Mathematics; Equations (Mathematics); Higher Education; Mathematical Applications; *Mathematical Concepts; Mathematics Education; *Proof (Mathematics); *Regression (Statistics); *Statistics; Supplementary Reading Materials

ABSTRACT

This is the third in a series of documents designed to supplement the statistics training of social science students studying applied statistics. The intent is to present selected proofs and derivations of important relationships or formulas that students typically do not find available and/or comprehensible in journals, textbooks, and similar sources. It is felt the unique feature of this material is the detailed step-by-step approach to all selected proofs and derivations. Calculus is neither assumed or used. Topics addressed include: (1) Introduction to Proof; (2) Derivation for Two Predictors; (3) Regression Model for Two Standardized Predictors; (4) Multiple Correlation for Two Standardized Predictors: (5) Derivation; (6) Normal Equations and Multiple Correlation Formula for Two Standardized Predictors; (7) Derivation for Three Predictors; (8) Normal Equations and Multiple Correlation Formula for Three Standardized Predictors; (9) Derivation for p Predictors; (10) Multiple Correlation for p Standardized Predictors and Derivation; and (11) Normal Equations and Multiple Correlation Formula for p Standardized Predictors. A single appendix provides details on Finding Normal Equations. (Author/MP)

A Derivation of the Sample Multiple Correlation Formula for Standard Scores

Francis J. O'Brien, Jr., Ph.D.

National Opinion Research Center

NORC
Sampling Department
461 8th Avenue
New York, New York 10001

November 22, 1982

Table of Contents

Page

Introduction 1

Introduction to Proof 1

Derivation for Two Predictors 2

Regression Model for Two Predictors in Standard Score Form 4

Multiple Correlation for Two Standardized Predictors 6

Derivation 8

Derivation for Three Predictors 11

Derivation for p Predictors 17

Regression Model for p Predictors in Standard Score Form 17

Multiple Correlation for p Standardized Predictors and Derivation 21

Appendix: Finding Normal Equations 26

Introduction 26

Plan 27

Finding Normal Equations for the Two Predictor Model 27

Finding Normal Equations for p Predictors 31

Alternate Procedure 33

Example for Five Predictors 34

Note 38

References 39

List of Tables

Table		Page
1.	Normal Equations and Multiple Correlation Formula for Two Standardized Predictors	9
2.	Normal Equations and Multiple Correlation Formula for Three Standardized Predictors	13
3.	Normal Equations and Multiple Correlation Formula for p Standardized Predictors	22

A Derivation of the Sample Multiple Correlation Formula for Standard Scores

Francis J. O'Brien, Jr., Ph.D.

This is the third paper in a series of publications that is designed to supplement the statistics training of students (see O'Brien, 1982a, 1982b). The intended audience is social science students studying applied statistics.

What I am attempting to do in this series is present selected proofs and derivations of important relationships or formulas that students do not find available and/or comprehensible in journals, textbooks and so forth. The unique feature of these papers is detailed step by step proofs or derivations. Calculus is not assumed nor is it used. Each proof or derivation is presented algebraically in great detail.

Introduction to Proof

Many students have learned that the multiple correlation between a criterion (or dependent variable) and a finite number of predictors (or independent variables) can be expressed as a weighted sum of regression weights and criterion/predictor product moment correlations. This relationship holds only for variables that are in standard score (z) form. This multiple correlation formula for p predictors (i.e., any number) can be written:

$$R_{Z_y \cdot Z_1, Z_2, \ldots, Z_j, \ldots, Z_p} = \sqrt{B_1 r_{y1} + B_2 r_{y2} + \ldots + B_j r_{yj} + \ldots + B_p r_{yp}}$$

Writing the right hand side in summation notation:

$$R_{Z_y \cdot Z_1, Z_2, \ldots, Z_j, \ldots, Z_p} = \sqrt{\sum_{j=1}^{p} B_j r_{yj}} \quad , \text{ where}$$

$R_{Z_y \cdot Z_1, Z_2, \ldots, Z_j, \ldots, Z_p}$ = multiple correlation of standardized variables

Z_y = the criterion expressed in standard score form

$Z_1, Z_2, \ldots, Z_j, \ldots, Z_p$ = standardized predictors

$B_1, B_2, \ldots, B_j, \ldots, B_p$ = beta (regression) weights attached to each standardized predictor

$r_{y1}, r_{y2}, \ldots, r_{yj}, \ldots r_{yp}$ = product moment criterion/predictor correlations.

The above multiple correlation formula is presented in many applied statistics textbooks. It will be derived in this paper. We will begin by deriving the relationship for the simplest multivariate case: one criterion and two predictors.

It is always helpful to have a plan in a proof or derivation. The general plan we will use can be summarized in the following steps:

1. state the regression model
2. derive the normal equations (See the Appendix)
3. define the multiple correlation
4. substitute the normal equations into the multiple correlation
5. simplify.

Some of these steps will be refined to suit a particular application.

Derivation for Two Predictors

Let us review briefly some of concepts, notation and logic in regression analysis. We will begin with regression analysis for two predictors in raw score form.

The mathematical function used to obtain the best linear fit for two raw score predictors is:

$\hat{Y} = a + b_1 X_1 + b_2 X_2$, where

$\hat{Y}$ = the predicted criterion

a, b_1 and b_2 = constants to be derived through the least squares procedure

X_1, X_2 = predictor variables

The notation in the above model is stated in abbreviated form. We have done this to minimize the reading of the symbolism and to clarify the concepts in the development of the derivation.

The regression model stated on page 2 is an idealized model. If a data set consisting of one criterion and two predictors can be assumed to be linear, then the model is a reasonable one to apply for prediction of actual or observed sample scores. It is idealized in the sense that no error term is included in the model. That is, when an actual or observed criterion score is compared to the criterion score predicted by the idealized model, some error is likely to occur—— the "fit" is less than perfect. If we call the actual sample score criterion Y, we can express the observed raw score model as follows:

$Y = \hat{Y} + e$, where

e = amount of numerical error resulting from using the idealized model ($\hat{Y}$) to predict the actual score (Y).

The goal in regression analysis is prediction of all individual criterion scores in a distribution with the smallest possible error. The error made in predicting observed criterion scores by the idealized model is:

$$Y - \hat{Y} = e$$

This is the quantity that we want to be as small as possible. The procedure most often used in the social sciences to accomplish this is the least squares procedure. The least squares criterion or goal can be expressed as:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \text{a minimum}^1$$

If we substitute the quantity for $\hat{Y}$ we can write:

$$\sum[Y - (a + b_1X_1 + B_2X_2)]^2 = \sum(Y - a - b_1X_1 - b_2X_2)^2 = \sum e^2 = \text{minimum}$$

---

[1] If it is understood that the all summations range from i=1 to i=n, then we can drop the summation limits all together; n, of course, refers to the sample size of n sample cases regardless of the number of variables in the regression model. Later when the algebra becomes more complex we use summation limits.

Regression Model for Two Standardized Predictors

If we now convert the variables of the two-predictor raw score model to standard score form, we can write the two predictor regression model as:

$$\frac{\hat{Y} - \bar{Y}}{S_{\hat{Y}}} = Z_{\hat{Y}} = A + B_1\left(\frac{X_1 - \bar{X}_1}{S_{X_1}}\right) + B_2\left(\frac{X_2 - \bar{X}_2}{S_{X_2}}\right), \text{ where}$$

$Z_{\hat{Y}}$ = the predicted criterion in standard score

A, B_1, B_2 = the constants to be derived through least squares

$\bar{X}_1, \bar{X}_2$ = sample means

S_{X_1}, S_{X_2} = sample standard deviations

We can write the regression model as:

$$Z_{\hat{Y}} = A + B_1Z_1 + B_2Z_2 \text{ , where}$$

Z_1, Z_2 = standard scores of the predictors.

The least squares criterion for the standard score regression model can be written as the difference between actual criterion and predicted criterion:

$$\sum(Z_Y - Z_{\hat{Y}})^2 = \sum e_Z^2 = \text{a minimum}$$

Substituting for $Z_{\hat{Y}}$,

$$\sum(Z_Y - A - B_1Z_1 - B_2Z_2)^2 = \sum e_Z^2 = \text{a minimum}$$

The least squares criterion is satisfied mathematically through calculus (partial derivatives). One by-product of the calculus technique is the so-called "normal equations". In the Appendix of this paper the reader will find a de-

scription of procedures given which will allow the reader to find normal equations for the two predictor model as well as models containing any number of predictors. The reader may find it helpful to study the Appendix at this point.

For the two standardized predictor model, the normal equations may be written as follows:

$$\sum z_Y = nA + B_1\sum z_1 + B_2\sum z_2$$
$$\sum z_Y z_1 = A\sum z_1 + B_1\sum z_1^2 + B_2\sum z_1 z_2$$
$$\sum z_Y z_2 = A\sum z_2 + B_1\sum z_1 z_2 + B_2\sum z_2^2$$

The normal equations can be simplified if the following facts about standard scores are recalled (see O'Brien, 1982b):

1. $\sum z_Y = \sum z_1 = \sum z_2 = 0$

2. $\sum z_1^2 = \sum z_2^2 = n-1$

3. a) $\frac{\sum z_y z_1}{n-1} = r_{y1}$ or $\sum z_y z_1 = (n-1)r_{y1}$

   b) $\frac{\sum z_y z_2}{n-1} = r_{y2}$ or $\sum z_y z_2 = (n-1)r_{y2}$

   c) $\frac{\sum z_1 z_2}{n-1} = r_{12}$ or $\sum z_1 z_2 = (n-1)r_{12}$

If these substitutions are made into the normal equations, we obtain:

$$0 = nA + 0 + 0$$
$$(n-1)r_{y1} = 0 + B_1(n-1) + B_2(n-1)r_{12}$$
$$(n-1)r_{y2} = 0 + B_1(n-1)r_{12} + B_2(n-1)$$

Consequently, A = 0 and may be ignored in subsequent results. If we divide through the last two normal equations by (n-1) we obtain a final statement of the normal equations:

$$r_{y1} = B_1 + B_2 r_{12}$$

$$r_{y2} = B_1 r_{12} + B_2$$

For the readers convenience we will restate these normal equations prior to the derivation.

Multiple Correlation for Two Standardized Predictors

By definition, the multiple correlation for two standardized predictors is:

$$R_{Z_Y \cdot Z_1, Z_2} = \text{corr}(Z_Y, Z_{\hat{Y}}) = \text{corr}(Z_Y, B_1 Z_1 + B_2 Z_2)$$

$$= \frac{\text{cov}(Z_Y, Z_{\hat{Y}})}{\sqrt{\text{var}(Z_Y)\text{var}(Z_{\hat{Y}})}}$$

$$= \frac{\text{cov}(Z_Y, B_1 Z_1 + B_2 Z_2)}{\sqrt{\text{var}(Z_Y)\text{var}(B_1 Z_1 + B_2 Z_2)}}, \text{ where}$$

corr means correlation; cov means covariance, and var means variance.

It is important to remember that B_1 and B_2 are constants. If we perform covariance and variance operations on the above correlation formula, we obtain:

$$R_{Z_Y \cdot Z_1, Z_2} = \frac{\text{cov}(Z_Y, B_1 Z_1) + \text{cov}(Z_Y, B_2 Z_2)}{\sqrt{\text{var}(B_1 Z_1) + \text{var}(B_2 Z_2) + 2\text{cov}(B_1 Z_1, B_2 Z_2)}}$$

Now, the $\sqrt{\text{var}(Z_Y)}$ is equal to 1 because Z_Y is a distribution of observed standardized criterion scores in a sample; that is,

$$\sqrt{\text{var}(Z_Y)} = \sqrt{\text{var}[(Y-\bar{Y})/S_Y]} = \sqrt{\text{var}(Y)/S_Y^2} = \sqrt{1} = 1.$$

The $\sqrt{\text{var}(Z_{\hat{Y}})}$, however, is the variance of predicted scores that comprise the regression plane of two variables; i.e., $\sqrt{\text{var}(B_1 Z_1 + B_2 Z_2)}$.

That is, $\sqrt{\text{var}(Z_{\hat{Y}})}$ is the variance of the equation used in Z_Y score prediction.

If we now apply rules of covariance and variance for standard score variables and constants, we can simplify the above correlation formula:

$$R_{Z_Y \cdot Z_1, Z_2} = \frac{B_1 \text{cov}(Z_Y, Z_1) + B_2 \text{cov}(Z_Y, Z_2)}{\sqrt{B_1^2 \text{var}(Z_1) + B_2^2 \text{var}(Z_2) + 2B_1 B_2 \text{cov}(Z_1, Z_2)}}$$

Further simplification can be achieved if it is recalled that:

1. the covariance of Z_Y and Z_1 is: $\text{cov}(Z_Y, Z_1) = r_{Z_Y Z_1} S_{Z_Y} S_{Z_1}$,

   but $r_{Z_Y Z_1} = r_{Y1}$ and $S_{Z_Y} = S_{Z_1} = 1$ (See O'Brien, 1982b);

   similar reasoning can be applied to $\text{cov}(Z_Y, Z_2)$ and $\text{cov}(Z_1, Z_2)$.

   Therefore, $\text{cov}(Z_Y, Z_1) = r_{y1}$
   $\text{cov}(Z_y, Z_2) = r_{y2}$
   $\text{cov}(Z_1, Z_2) = r_{12}$

2. $\text{var}(Z_1) = \text{var}(Z_2) = 1$

Making these substitutions,

$$R_{Z_Y \cdot Z_1, Z_2} = \frac{B_1 r_{y1} + B_2 r_{y2}}{\sqrt{B_1^2 + B_2^2 + 2B_1 B_2 r_{12}}}$$

Derivation

We are now able to show the derivation of the following multiple correlation formula which appears in many applied statistics textbooks without proof:

$$R_{Z_Y \cdot Z_1, Z_2} = \sqrt{B_1 r_{y1} + B_2 r_{y2}}$$

For the readers convenience, we will restate the normal equations and the multiple correlation formula presented earlier. See Table 1.

The derivation consists of a) substituting the normal equations into the numerator of the multiple correlation formula and b) simplifying algebraically.

See the page following Table 1.

Table 1

Normal Equations and Multiple Correlation Formula for Two Standardized Predictors

---

Normal Equations

$$r_{y1} = B_1 + B_2 r_{12}$$

$$r_{y2} = B_1 r_{12} + B_2$$

Multiple Correlation Formula

$$R_{Z_Y \cdot Z_1, Z_2} = \frac{B_1 r_{y1} + B_2 r_{y2}}{\sqrt{B_1^2 + B_2^2 + 2B_1 B_2 r_{12}}}$$

Note: Proof that $R_{Z_Y \cdot Z_1, Z_2} = \sqrt{B_1 r_{y1} + B_2 r_{y2}}$ requires substituting the normal equations into the numerator of the multiple correlation formula and simplifying. See text for details.

If we substitute the normal equations for r_{y1} and r_{y2} into the numerator of the multiple correlation formula we obtain (see Table 1):

$$R_{Z_Y \cdot Z_1, Z_2} = \frac{B_1(B_1 + B_2 r_{12}) + B_2(B_1 r_{12} + B_2)}{\sqrt{B_1^2 + B_2^2 + 2B_1 B_2 r_{12}}}$$

$$= \frac{B_1^2 + B_1 B_2 r_{12} + B_1 B_2 r_{12} + B_2^2}{\sqrt{B_1^2 + B_2^2 + 2B_1 B_2 r_{12}}}$$

$$= \frac{B_1^2 + B_2^2 + 2B_1 B_2 r_{12}}{\sqrt{B_1^2 + B_2^2 + 2B_1 B_2 r_{12}}}$$

(Thus, $B_1 r_{y1} + B_2 r_{y2} = B_1^2 + B_2^2 + 2B_1 B_2 r_{12}$)

$$= \sqrt{B_1^2 + B_2^2 + 2B_1 B_2 r_{12}}$$

$$= \sqrt{B_1 r_{y1} + B_2 r_{y2}}$$

END OF PROOF[1]

[1]Recall from algebra that for any algebraic term:

$$\frac{A}{\sqrt{A}} = \frac{A\left(\sqrt{A}\right)}{\sqrt{A}\left(\sqrt{A}\right)} = \frac{A\sqrt{A}}{\sqrt{A^2}} = \frac{A\sqrt{A}}{A} = \sqrt{A}$$

Derivation for Three Predictors

Prior to the derivation for p predictors (the general case), let us consider the case for 3 predictors. This will allow us to review the logic of the derivation. In addition, we will introduce the use of summation notation which is necessary to do for the p predictor derivation.

The first step is to state the regression model for three standardized predictors:

$$Z_{\hat{Y}} = A + B_1Z_1 + B_2Z_2 + B_3Z_3$$

The least squares criterion for this model is:

$$\sum(Z_Y - Z_{\hat{Y}})^2 = \sum[Z_Y - A - B_1Z_1 - B_2Z_2 - B_3Z_3]^2 = \text{a minimum}$$

The next step is to derive the normal equations. As outlined in the Appendix, the normal equations for 3 predictors (in simplified form) are:

$$r_{y1} = B_1 + B_2r_{12} + B_3r_{13}$$

$$r_{y2} = B_1r_{12} + B_2 + B_3r_{23}$$

$$r_{y3} = B_1r_{13} + B_2r_{23} + B_3$$

For the readers convenience we will restate the normal equations prior to the derivation.

The third step is to define the multiple correlation between Z_Y and the three predictors. This is done on the following page.

$$R_{Z_Y \cdot Z_1,Z_2,Z_3} = \text{corr}(Z_Y, Z_{\hat{Y}}) = \text{corr}(Z_Y, B_1Z_1 + B_2Z_2 + B_3Z_3)$$

$$= \frac{\text{cov}(Z_Y,\ Z_{\hat{Y}})}{\sqrt{\text{var}(Z_Y)\text{var}(Z_{\hat{Y}})}}$$

$$= \frac{\text{cov}(Z_Y,\ B_1Z_1 + B_2Z_2 + B_3Z_3)}{\sqrt{\text{var}(B_1Z_1 + B_2Z_2 + B_3Z_3)}}$$

Recall that $\sqrt{\text{var}(Z_Y)} = 1$.

Applying the rules of covariance and variance algebra for standard score variables and constants:

$$R_{Z_Y \cdot Z_1,Z_2,Z_3} = \frac{\text{cov}(Z_Y,B_1Z_1) + \text{cov}(Z_Y,B_2Z_2) + \text{cov}(Z_Y,B_3Z_3)}{\sqrt{\text{var}(B_1Z_1) + \text{var}(B_2Z_2) + \text{var}(B_3Z_3) + 2\text{cov}(B_1Z_1,B_2Z_2) + 2\text{cov}(B_1Z_1,B_3Z_3) + 2\text{cov}(B_2Z_2,\ B_3Z_3)}}$$

$$= \frac{B_1r_{y1} + B_2r_{y2} + B_3r_{y3}}{\sqrt{B_1^2\text{var}(Z_1) + B_2^2\text{var}(Z_2) + B_3^2\text{var}(Z_3) + 2B_1B_2\text{cov}(Z_1,Z_2) + 2B_1B_3\text{cov}(Z_1,Z_3) + 2B_2B_3\text{cov}(Z_2,Z_3)}}$$

$$= \frac{B_1r_{y1} + B_2r_{y2} + B_3r_{y3}}{\sqrt{B_1^2 + B_2^2 + B_3^2 + 2B_1B_2r_{12} + 2B_1B_3r_{13} + 2B_2B_3r_{23}}}$$

For easy reference, the multiple correlation and normal equations are restated in Table 2.

Table 2

Normal Equations and Multiple Correlation Formula for Three Standardized Predictors

---

Normal Equations

$$r_{y1} = B_1 + B_2 r_{12} + B_3 r_{13}$$

$$r_{y2} = B_1 r_{12} + B_2 + B_3 r_{23}$$

$$r_{y3} = B_1 r_{13} + B_2 r_{23} + B_3$$

Multiple Correlation Formula

$$R_{Z_Y \cdot Z_1, Z_2, Z_3} = \frac{B_1 r_{y1} + B_2 r_{y2} + B_3 r_{y3}}{\sqrt{B_1^2 + B_2^2 + B_3^2 + 2B_1 B_2 r_{12} + 2B_1 B_3 r_{13} + 2B_2 B_3 r_{23}}}$$

Note: Proof that $R_{Z_Y \cdot Z_1, Z_2, Z_3} = \sqrt{B_1 r_{y1} + B_2 r_{y2} + B_3 r_{y3}}$ requires substituting the normal equations into the numerator of the multiple correlation formula and simplifying. See text for details.

The fourth step is to substitute the normal equations into the numerator of the multiple correlation formula. Substituting the normal equations:

$$\text{cov}(Z_Y, Z_{\hat{Y}}) = B_1 r_{y1} + B_2 r_{y2} + B_3 r_{y3}$$

$$= B_1(B_1 + B_2 r_{12} + B_3 r_{13}) + B_2(B_1 r_{12} + B_2 + B_3 r_{23}) + B_3(B_1 r_{13} + B_2 r_{23} + B_3)$$

$$= (B_1^2 + B_1 B_2 r_{12} + B_1 B_3 r_{13}) + (B_1 B_2 r_{12} + B_2^2 + B_2 B_3 r_{23}) + (B_1 B_3 r_{13} + B_2 B_3 r_{23} + B_3^2)$$

If we write each parenthesized term on a separate line, we obtain:

$$\begin{aligned}\text{cov}(Z_Y, Z_{\hat{Y}}) = {} & B_1^2 + B_1 B_2 r_{12} + B_1 B_3 r_{13} + \\ & B_1 B_2 r_{12} + B_2^2 + B_2 B_3 r_{23} + \\ & B_1 B_3 r_{13} + B_2 B_3 r_{23} + B_3^2\end{aligned}$$

Let us pause for a moment and consider how to write out this covariance matrix in summation notation.

It is clear that the three squared B terms can be written in summation notation as:

$$B_1^2 + B_2^2 + B^2{}_3 = \sum_{j=1}^{3} B_j^2$$

The remainder of the terms consists of three pairs of quantities:

$$2B_1 B_2 r_{12} + 2B_1 B_3 r_{13} + 2B_2 B_3 r_{23}$$

One common way to write this in summation notation is as follows:

$$2(B_1 B_2 r_{12} + B_1 B_3 r_{13} + B_2 B_3 r_{23}) = 2 \sum_{j=2}^{3} \sum_{i=1}^{2} B_i B_j r_{ij}$$

The total number of terms can be determined by multiplying the upper limits of the summation (3x2 = 6). Also, from the summation limits(i=1,2 j=2,3) it is clear that the first term is $B_1 B_2 r_{12}$ and the last term is $B_2 B_3 r_{23}$. That

leaves 4 terms to be filled in. Start from $B_1B_2r_{12}$ and increment the summation limits by one— begin with i=1 and increment j until it is exhausted (2,3) and then go back to i and increment it to 2, and increment j until the limits are exhausted. It is helpful to have a covariance matrix written out such that the term pairs can be "read off".

Thus, the covariance term of the multiple correlation formula can be written in summation notation as follows:

$$\text{cov}(Z_Y, Z_{\hat{Y}}) = B_1^2 + B_2^2 + B_3^2 + 2B_1B_2r_{12} + 2B_1B_3r_{13} + 2B_2B_3r_{23}$$

$$= \sum_{j=1}^{3} B_j^2 + 2\sum_{j=2}^{3}\sum_{i=1}^{2} B_iB_jr_{ij}$$

$$= B_1r_{y1} + B_2r_{y2} + B_3r_{y3} = \sum_{j=1}^{3} B_jr_{yj}$$

Turning to the denominator of the correlation formula, it is readily apparent that it is identical to the covariance term above. That is:

$$\sqrt{\text{var}(Z_{\hat{Y}})} = \sqrt{B_1^2 + B_2^2 + B_3^2 + 2B_1B_2r_{12} + 2B_1B_3r_{13} + 2B_2B_3r_{23}}$$

$$= \sqrt{\sum_{j=1}^{3} B_j^2 + 2\sum_{j=2}^{3}\sum_{i=1}^{2} B_iB_jr_{ij}}$$

$$= \sqrt{\sum_{j=1}^{3} B_jr_{yj}}$$

Hence, the multiple correlation written in summation notation is:

$$R_{Z_Y \cdot Z_1, Z_2, Z_3} = \frac{\sum_{j=1}^{3} B_j^2 + 2\sum_{j=2}^{3}\sum_{i=1}^{2} B_iB_jr_{ij}}{\sqrt{\sum_{j=1}^{3} B_j^2 + 2\sum_{j=2}^{3}\sum_{i=1}^{2} B_iB_jr_{ij}}}$$

Making the same algebraic simplification we made for the two predictor derivation, we obtain:

$$R_{Z_Y \cdot Z_1, Z_2, Z_3} = \sqrt{\sum_{j=1}^{3} B_j^2 + 2\sum_{j=2}^{3}\sum_{i=1}^{2} B_i B_j r_{ij}}$$

$$= \sqrt{\sum_{j=1}^{3} B_j r_{yj}}$$

This completes the derivation for three predictors. We now derive the case for any number of predictors in the regression model.

Derivation for p Predictors

Regression Model for p Predictors in Standard Score Form

The derivation for any possible number of predictors (p) will be worked out following the same steps used in the derivation for 2 and 3 predictors. A restatement of these steps for p predictors is:

1. state the regression model for p predictors
2. derive the normal equations
3. define the multiple correlation
4. substitute the normal equations into the numerator of the the multiple correlation formula
5. express the covariance term in summation notation
6. express the variance term in summation notation
7. simplify algebraically

The linear regression model for p predictors in standard score form is:

$$Z_{\hat{Y}} = A + B_1Z_1 + B_2Z_2 + B_3Z_3 + \ldots + B_jZ_j + \ldots + B_pZ_p$$

(Shortly we will see that the A term is equal to 0 and can be ignored as we discovered in the 2 and 3 predictor models).

The least squares criterion is:

$$\sum(Z_Y - Z_{\hat{Y}})^2 = \sum e_Z^2 = \text{a minimum.}$$

Substituting for $Z_{\hat{Y}}$, the least squares criterion is:

$$\sum(Z_Y - A - B_1Z_1 - B_2Z_2 - B_3Z_3 - \ldots - B_jZ_j - \ldots - B_pZ_p)^2 = \text{minimum}$$

Finding the normal equations involves the same procedure used for the 2 or 3 predictor models. See the Appendix for details. The normal equations (before simplification) are stated on the following page.[1]

---

[1] Note that the equations for $\sum z_y z_1$, $\sum z_y z_2$ etc. are written with the subscripts reversed. Since these sums are symmetric such that $\sum z_1 z_2 = \sum z_2 z_1$, $\sum z_4 z_3 = \sum z_3 z_4$, or in general, $\sum z_i z_j = \sum z_j z_i$, we have written these terms such that the first subscript is always less than the second subscript. This method of notation helps simplify the algebra.

19.

$$\sum z_Y = nA + B_1\sum z_1 + B_2\sum z_2 + B_3\sum z_3 + \ldots + B_j\sum z_j + \ldots + B_p\sum z_p$$

$$\sum z_Y z_1 = A\sum z_1 + B_1\sum z_1^2 + B_2\sum z_1 z_2 + B_3\sum z_1 z_3 + \ldots + B_j\sum z_1 z_j + \ldots + B_p\sum z_1 z_p$$

$$\sum z_Y z_2 = A\sum z_2 + B_1\sum z_1 z_2 + B_2\sum z_2^2 + B_3\sum z_2 z_3 + \ldots + B_j\sum z_2 z_j + \ldots + B_p\sum z_2 z_p$$

$$\sum z_Y z_3 = A\sum z_3 + B_1\sum z_1 z_3 + B_2\sum z_2 z_3 + B_3\sum z_3^2 + \ldots + B_j\sum z_3 z_j + \ldots + B_p\sum z_3 z_p$$

$$\vdots$$

$$\sum z_Y z_p = A\sum z_p + B_1\sum z_1 z_p + B_2\sum z_2 z_p + B_3\sum z_3 z_p + \ldots + B_j\sum z_j z_p + \ldots + B_p\sum z_p^2$$

If we apply the same logic and make the same substitutions as we did previously for the two and three predictor models, we obtain a simplified set of normal equations:

$$r_{y1} = B_1 + B_2 r_{12} + B_3 r_{13} + \ldots + B_j r_{1j} + \ldots + B_p r_{1p}$$

$$r_{y2} = B_1 r_{12} + B_2 + B_3 r_{23} + \ldots + B_j r_{2j} + \ldots + B_p r_{2p}$$

$$r_{y3} = B_1 r_{13} + B_2 r_{23} + B_3 + \ldots + B_j r_{3j} + \ldots + B_p r_{3p}$$

$$\vdots$$

$$r_{yp} = B_1 r_{1p} + B_2 r_{2p} + B_3 r_{3p} + \ldots + B_j r_{jp} + \ldots + B_p$$

24

25

These are the normal equations we want to work with in the derivation. A restatement of them is given later.

To give the reader a "feel" for the notation in multivariate problems, we will work out the normal equations for 5 predictors.[1]

The least squares criterion is:

$$\sum(Z_Y - A - B_1Z_1 - B_2Z_2 - B_3Z_3 - B_4Z_4 - B_5Z_5)^2 = \text{a minimum}$$

Using the procedure outlined in the Appendix, the normal equations are as follows:

$$\sum Z_Y = nA + B_1\sum Z_1 + B_2\sum Z_2 + B_3\sum Z_3 + B_4\sum Z_4 + B_5\sum Z_5$$

$$\sum Z_YZ_1 = A\sum Z_1 + B_1\sum Z_1^2 + B_2\sum Z_1Z_2 + B_3\sum Z_1Z_3 + B_4\sum Z_1Z_4 + B_5\sum Z_1Z_5$$

$$\sum Z_YZ_2 = A\sum Z_2 + B_1\sum Z_1Z_2 + B_2\sum Z_2^2 + B_3\sum Z_2Z_3 + B_4\sum Z_2Z_4 + B_5\sum Z_2Z_5$$

$$\sum Z_YZ_3 = A\sum Z_3 + B_1\sum Z_1Z_3 + B_2\sum Z_2Z_3 + B_3\sum Z_3^2 + B_4\sum Z_3Z_4 + B_5\sum Z_3Z_5$$

$$\sum Z_YZ_4 = A\sum Z_4 + B_1\sum Z_1Z_4 + B_2\sum Z_2Z_4 + B_3\sum Z_3Z_4 + B_4\sum Z_4^2 + B_5\sum Z_4Z_5$$

$$\sum Z_YZ_5 = A\sum Z_5 + B_1\sum Z_1Z_5 + B_2\sum Z_2Z_5 + B_3\sum Z_3Z_5 + B_4\sum Z_4Z_5 + B_5\sum Z_5^2$$

Simplifying, we get:

$$r_{y1} = B_1 + B_2r_{12} + B_3r_{13} + B_4r_{14} + B_5r_{15}$$

$$r_{y2} = B_1r_{12} + B_2 + B_3r_{23} + B_4r_{24} + B_5r_{25}$$

$$r_{y3} = B_1r_{13} + B_2r_{23} + B_3 + B_4r_{34} + B_5r_{35}$$

$$r_{y4} = B_1r_{14} + B_2r_{24} + B_3r_{34} + B_4 + B_5r_{45}$$

$$r_{y5} = B_1r_{15} + B_2r_{25} + B_3r_{35} + B_4r_{45} + B_5$$

---

[1] The reader who has studied the Appendix may wish to attempt writing out the normal equations in advance of seeing the results.

Multiple Correlation for p Standardized Predictors and Derivation

We are now ready to derive the multiple correlation for p predictors. See Table 3 for a statement of the multiple correlation formula.

The covariance term is:

$$\text{cov}(Z_Y, B_1Z_1, B_2Z_2, B_3Z_3, \ldots, B_jZ_j, \ldots, B_pZ_p)$$

$$= B_1r_{y1} + B_2r_{y2} + B_3r_{y3} + \ldots + B_jr_{yj} + \ldots + B_pr_{yp}$$

Substituting the normal equations (see Table 3),

$$\text{cov}(Z_Y, Z_{\hat{Y}}) = B_1(B_1 + B_2r_{12} + \ldots + B_pr_{1p}) + B_2(B_1r_{12} + B_2 + \ldots + B_p$$

$$B_3(B_1r_{13} + B_2r_{23} + \ldots + B_pr_{3p}) + \ldots +$$

$$B_p(B_1r_{1p} + B_2r_{2p} + \ldots + B_p)$$

In order to express this in summation notation, let us write out the full $\text{cov}(Z_Y, Z_{\hat{Y}})$. Multiply each B_j term inside the parentheses and state each parenthesized term on a separate line, giving:

$$\text{cov}(Z_Y, Z_{\hat{Y}}) = B_1r_{y1} + B_2r_{y2} + B_3r_{y3} + \ldots + B_jr_{yj} + \ldots + B_pr_{yp} =$$

$$B_1^2 + B_1B_2r_{12} + B_1B_3r_{13} + \ldots + B_1B_jr_{1j} + \ldots + B_1B_pr_{1p}$$

$$B_1B_2r_{12} + B_2^2 + B_2B_3r_{23} + \ldots + B_2B_jr_{2j} + \ldots + B_2B_pr_{2p}$$

$$B_1B_3r_{13} + B_2B_3r_{23} + B_3^2 + \ldots + B_3B_jr_{3j} + \ldots + B_3B_pr_{3p}$$

$$\vdots$$

$$B_1B_pr_{1p} + B_2B_pr_{2p} + B_3B_pr_{3p} + \ldots + B_jB_pr_{jp} + \ldots + B_p^2$$

Table 3

Normal Equations and Multiple Correlation Formula for p Standardized Predictors

22.

Normal Equations

$$r_{y1} = B_1 + B_2 r_{12} + B_3 r_{13} + \ldots + B_j r_{1j} + \ldots + B_p r_{1p}$$

$$r_{y2} = B_1 r_{12} + B_2 + B_3 r_{23} + \ldots + B_j r_{2j} + \ldots + B_p r_{2p}$$

$$r_{y3} = B_1 r_{13} + B_2 r_{23} + B_3 + \ldots + B_j r_{3j} + \ldots + B_p r_{3p}$$

$$\vdots$$

$$r_{yp} = B_1 r_{1p} + B_1 r_{2p} + B_3 r_{3p} + \ldots + B_j r_{jp} + \ldots + B_p$$

Multiple Correlation Formula

$$R_{z_Y \cdot z_1, z_2, z_3, \ldots, z_j, \ldots, z_p} = \text{corr}(Z_Y, Z_{\hat{Y}}) = \text{corr}(Z_Y, B_1 Z_1, B_2 Z_2, B_3 Z_3, \ldots, B_j Z_j, \ldots, B_p Z_p)$$

$$= \frac{\text{cov}(Z_Y, Z_{\hat{Y}})}{\sqrt{\text{var}(Z_Y)\ \text{var}(Z_{\hat{Y}})}}$$

$$= \frac{\text{cov}(Z_Y, B_1 Z_1 + B_2 Z_2 + B_3 Z_3 + \ldots + B_j Z_j + \ldots + B_p Z_p)}{\sqrt{\text{var}(B_1 Z_1 + B_2 Z_2 + B_3 Z_3 + \ldots + B_j Z_j + \ldots + B_p Z_p)}}$$

$$= \frac{B_1 r_{y1} + B_2 r_{y2} + B_3 r_{y3} + \ldots B_j r_{yj} + \ldots + B_p r_{yp}}{\sqrt{B_1^2 + B_2^2 + B_3^2 + \ldots + B_j^2 + \ldots + B_p^2 + 2B_1 B_2 r_{12} + 2B_1 B_3 r_{13} + 2B_1 B_4 r_{14} + \ldots + 2B_1 B_j r_{ij} + \ldots + 2 B_{p-1} B_p r_{p-1,p}}}$$

27

Note: proof requires substituting the normal equations into the numerator of the multiple correlation formula and simplifying. See text for details.

23

We want to express this sum in summation notation to simplify the algebra. First, we count the number of terms to be summed. It is evident that each row of the covariance matrix consists of p terms. Also, there are a total of p rows. Hence, there are a total of $(p)(p) = p^2$ terms in the entire matrix. It is also evident that each row contain one B_j^2 term or a total of p B_j^2 terms in the entire matrix (along the northwest to southeast diagonal of the matrix). How many other terms are in the matrix (off diagonal terms) can be answered with a little algebra. Let X represent the number of off diagonal ($B_i B_j r_{ij}$) terms. Then:

$$p^2 = p + X \text{ or}$$

$$X = p^2 - p$$

$$X = p(p-1)$$

Thus, there are p B_j^2 terms and $p(p-1)$ $B_i B_j r_{ij}$ terms in the entire matrix.

Another view is as follows. The diagonal (B_j^2 terms) consists of p terms. The remainder of the off diagonal terms consists of a number of identical pairs of terms. If we halve the matrix and visualize the upper half only, then we are thinking of p B_j^2 terms plus one-half of the $B_i B_j r_{ij}$ terms. That is, because of the symmetry of the off diagonal terms in the matrix, the off diagonal consists of $\frac{p(p-1)}{2}$ terms. The total number of terms in the half matrix is: $p + \frac{p(p-1)}{2}$. To represent the total number of terms in the entire matrix, simply double the number of off diagonal terms in the half matrix: $p^2 = p + 2[p(p-1)/2] = p + p(p-1)$. Examine the matrix of 5 predictors on page 20 for clarification.

As explained, the $\text{cov}(Z_Y, Z_Y)$ consists of p^2 terms. There are p B_j^2 and $p(p-1)$ $B_i B_j r_{ij}$ (or $2[p(p-1)/2]$) terms. Consequently, we can write

the covariance of the multiple correlation formula as:

$$\text{cov}(Z_Y, Z_{\hat{Y}}) = \sum_{j=1}^{p} B_j^2 + 2\sum_{j=2}^{p}\sum_{i=1}^{p-1} B_i B_j r_{ij}$$

This is simply a generalization of the 2 and 3 predictor models. The first term of the double summation will be $B_1 B_2 r_{12}$ and the last term will be $B_{p-1} B_p r_{p-1,p}$ (see Table 3). For example, in a 5 predictor model, there are $5^2 = 25$ terms; 5 are B_j^2 terms and $5(5-1) = 20$ are $B_i B_j r_{ij}$ terms (or 10 pairs of terms [$2p(p-1)/2 = 2 \times 5(4)/2 = 2 \times 10 = 20$]). The first term will be $B_1 B_2 r_{12}$ and the last will be $B_4 B_5 r_{45}$.

Therefore we can express the covariance term in the p predictor model as:

$$\text{cov}(Z_Y, Z_{\hat{Y}}) = B_1 r_{y1} + B_2 r_{y2} + B_3 r_{y3} + \ldots + B_j r_{yj} + \ldots + B_p r_{yp}$$

$$= \sum_{j=1}^{p} B_j r_{yj}$$

Equivalently,

$$\text{cov}(Z_Y, Z_{\hat{Y}}) = \sum_{j=1}^{p} B_j^2 + 2\sum_{j=2}^{p}\sum_{i=1}^{p-1} B_i B_j r_{ij}$$

Thus,

$$\text{cov}(Z_Y, Z_{\hat{Y}}) = \sum_{j=1}^{p} B_j r_{yj} = \sum_{j=1}^{p} B_j^2 + 2\sum_{j=2}^{p}\sum_{i=1}^{p-1} B_i B_j r_{ij}$$

Turning now to the variance term of the correlation formula, we can apply rules of variance and covariance algebra to the $B_j Z_j$ terms of $Z_{\hat{Y}}$ (see Table 3). Again, the results of these manipulations are simply generalizations of the 2 and 3 predictor models. From an inspection of Table 3, it should be apparent that we can express the variance term as follows:

$$\sqrt{var(Z_{\hat{Y}})} = \sqrt{var(B_1Z_1 + B_2Z_2 + B_3Z_3 + \ldots + B_jZ_j + \ldots B_pZ_p)}$$

$$= \sqrt{\begin{array}{l} B_1^2 + B_2^2 + B_3^2 + \ldots + B_j^2 + \ldots + B_p^2 + \\ 2B_1B_2r_{12} + 2B_1B_3r_{13} + 2B_1B_4r_{14} + \ldots + 2B_iB_jr_{ij} + \ldots + 2B_{p-1}B_pr_{p-1,p} \end{array}}$$

Thus, in summation notation:

$$\sqrt{var(Z_{\hat{Y}})} = \sqrt{\sum_{j=1}^{p} B_j^2 + 2\sum_{j=2}^{p}\sum_{i=1}^{p-1} B_iB_jr_{ij}}$$

Therefore, the multiple correlation for p predictors is:

$$R_{Z_Y \cdot Z_1,Z_2,Z_3,\ldots,Z_j,\ldots,Z_p} = \frac{\sum_{j=1}^{p} B_j^2 + 2\sum_{j=2}^{p}\sum_{i=1}^{p-1} B_iB_jr_{ij}}{\sqrt{\sum_{j=1}^{p} B_j^2 + 2\sum_{j=2}^{p}\sum_{i=1}^{p-1} B_iB_jr_{ij}}}$$

$$= \sqrt{\sum_{j=1}^{p} B_j^2 + 2\sum_{j=2}^{p}\sum_{i=1}^{p-1} B_iB_jr_{ij}}$$

$$= \sqrt{\sum_{j=1}^{p} B_jr_{yj}}$$

END OF PROOF

APPENDIX

Finding Normal Equations

Introduction

This appendix will outline in detail the procedures for finding the normal equations in regression analysis. The procedures described are appropriate when:

a) the regression model is linear and

b) the variables are standardized.

An example of a nonlinear model is:

$$Z_{\hat{Y}} = A + B_1 Z_1 + B_2 Z_2^2$$

The model is not linear because one of the variables (Z_2) is raised to the second power (Z_2^2). The procedures described in this appendix would not be applicable for such a model.

Just what is a normal equation? This is a question often asked by students. One way to answer this question is to say that a normal equation is one of the equations that results from a calculus technique called partial differentiation applied to a regression model to satisfy a criterion of minimization. For example, a regression model of two predictors contains three constants (A, B_1, B_2) which must be solved in order to minimize the function $\sum(Z_Y - Z_{\hat{Y}})^2 = \sum(Z_Y - A - B_1 Z_1 - B_2 Z_2)^2 = \sum e_Z^2$ using the least squares criterion (smallest or minimum error when the idealized model is used for prediction of actual or observed criterion scores, Z_Y). In this example, the calculus procedure is applied to each of the three terms individually. When the procedure is applied for one of the terms, and the result is solved algebraically in terms of the criterion variable, the result is termed the "normal equation" for that term.

In this paper we are not interested in solving for the terms of the model per se. Rather we are interested in using the normal equations to derive the multiple correlation formula for standardized scores. The normal equations allow one to accomplish this. In fact, we are following the same steps that would be used in actually solving for the terms that satisfy the least squares criterion up to the point when the normal equations are derived for a regression model. Since this paper does not assume a knowledge of calculus, a heuristic procedure is given for finding normal equations. Students who are familar with calculus can read any text of mathematical statistics for technical details.

Plan

The plan for finding normal equations is outlined in four phases as follows:

A. state the regression model, $Z_{\hat{Y}}$

B. state the mathematical function of the least squares criterion, $\sum(Z_Y - Z_{\hat{Y}})^2$

C. derive the normal equations for each of the terms and simplify

D summarize the normal equations

Finding Normal Equations for the Two Predictor Model

We will demonstrate the four phase procedure first for the 2 predictor model.

A. The regression model for 2 predictors:

$$Z_{\hat{Y}} = A + B_1Z_1 + B_2Z_2$$

B. The mathematical function to be minimized according to the least squares criterion is:

$$\sum(Z_Y - Z_{\hat{Y}})^2 = \sum(Z_Y - A - B_1Z_1 - B_2Z_2)^2$$

C. The procedures for deriving the normal equations for the constants are as follows:

1. The procedure for finding the normal equation for A is summarized in 5 steps:
   1. drop exponent and set function in phase B equal to 0 .
   2. distribute the summation operator
   3. apply the rules of summation for constants
   4. solve in terms of the criterion variable, Z_Y
   5. Apply the rules for standard scores and simplify.[1]

Applying each of the steps in turn produces:

1. $\sum(Z_Y - A - B_1Z_1 - B_2Z_2) = 0$
2. $\sum Z_Y - \sum A - \sum B_1Z_1 - \sum B_2Z_2 = 0$
3. $\sum Z_Y - nA - B_1\sum Z_1 - B_2\sum Z_2 = 0$
4. $\sum Z_Y = nA + B_1\sum Z_1 + B_2\sum Z_2$
5. $0 = nA + B_1 0 + B_2 0$

Solving for A in step 5 shows that A = 0. As a general rule, A will always be 0 when the regression function is linear and stated in standard score form.

---

[1] For students who need to review standard scores, see O'Brien, 1982b. Also, see page 5 of the present paper for examples of rules.

2. The procedure for finding the normal equation for B_1 can be summarized in 7 steps:

   1. drop the exponent 2 and set function in phase B equal to 0
   2. multiply the function by Z_1
   3. distribute the Z_1 term
   4. distribute the summation operator
   5. apply rules of summation for constants
   6. solve in terms of the criterion variable, Z_Y
   7. apply rules for standard scores and simplify.

Note that we do not try to solve for B_1 in this procedure. We are applying a procedure to find the normal equation for B_1. Applying the above 7 steps:

1. $\sum(Z_Y - A - B_1Z_1 - B_2Z_2) = 0$
2. $\sum(Z_Y - A \quad B_1Z_1 - B_2Z_2)Z_1 = 0$
3. $\sum(Z_YZ_1 - AZ_1 - B_1Z_1^2 - B_2Z_1Z_2) = 0$
4. $\sum Z_YZ_1 - \sum AZ_1 - \sum B_1Z_1^2 - \sum B_1Z_1Z_2 = 0$
5. $\sum Z_YZ_1 - A\sum Z_1 - B_1\sum Z_1^2 - B_2\sum Z_1Z_2 = 0$
6. $\sum Z_YZ_1 = A\sum Z_1 - B_1\sum Z_1^2 - B_2\sum Z_1Z_2$
7. $(n-1)r_{y1} = 0 + B_1(n-1) + B_2(n-1)r_{12}$

Dividing through by n-1, we obtain a simplified statement of the normal equation for B_1:

$$r_{y1} = B_1 + B_2r_{12}$$

3. The steps for finding the normal equation for B_2 parallel those for B_1:

   1. drop the exponent 2 and set function in phase B equal to 0
   2. multiply the function by Z_2
   3. distribute the Z_2 term
   4. distribute the summation operator
   5. apply rules of summation for constants
   6. solve in terms of the criterion variable
   7. apply rules for standard scores and simplify.

   To iterate, we are not solving for B_2. Applying the 7 rules:

   1. $\sum(Z_Y - A - B_1Z_1 - B_2Z_2) = 0$
   2. $\sum(Z_Y - A - B_1Z_1 - B_2Z_2)Z_2 = 0$
   3. $\sum(Z_YZ_2 - AZ_2 - B_1Z_1Z_2 - B_2Z_2^2) = 0$
   4. $\sum Z_YZ_2 - \sum AZ_2 - \sum B_1Z_1Z_2 - \sum B_2Z_2^2 = 0$
   5. $\sum Z_YZ_2 - A\sum Z_2 - B_1\sum Z_1Z_2 - B_2\sum Z_2^2 = 0$
   6. $\sum Z_YZ_2 = A\sum Z_2 + B_1\sum Z_1Z_2 + B_2\sum Z_2^2$
   7. $(n-1)r_{y2} = 0 + B_1(n-1)r_{12} + B_2(n-1)$

   Dividing through by $(n-1)$,

$$r_{y2} = B_1r_{12} + B_2$$

D. We now write out a full statement of the normal equations to summarize the results. As noted, a normal equation is considered derived at the point when we solve in terms of the criterion variable. Subsequent steps are employed to simplify the result. The normal equations for A, B_1 and B_2 were:

$$\text{for A: } \sum z_y = nA + B_1\sum z_1 + B_2\sum z_2$$

$$\text{for } B_1\text{: } \sum z_y z_1 = A\sum z_1 + B_1\sum z_1^2 + B_1\sum z_1 z_2$$

$$\text{for } B_2\text{: } \sum z_y z_2 = A\sum z_2 + B_1\sum z_1 z_2 + B_2\sum z_2^2$$

When we applied rules for standard scores and simplified, we obtained the following set of normal equations used in the derivation for two predictors (recall A=0):

$$r_{y1} = B_1 + B_2 r_{12}$$

$$r_{y2} = B_1 r_{12} + B_2$$

Finding Normal Equations for p Predictors

The rules for finding normal equations for the two predictor model can be generalized readily for models with any number (p) of predictors. We will show two methods for the general case. First we will demonstrate the procedure using the four phase plan. Then we show a much simpler method. But the shorter method depends on first showing the longer one.

Applying the four phase plan gives the following results:

A. The regression model for p predictors is:

$$Z_{\hat{Y}} = A + B_1Z_1 + {}_2Z_2 + B_3Z_3 +\ldots+ B_jZ_j +\ldots+ B_pZ_p$$

B. The function to be minimized according to the least squares criterion is:

$$\sum(Z_Y - A - B_1Z_1 - B_2Z_2 - B_3Z_3 -\ldots- B_jZ_j -\ldots- B_pZ_p)^2$$

C. The procedures for finding normal equations for A and <u>any</u> B_j term are:

1. In deriving the normal equation for A, the result is always the same— $A = 0$.

2. Finding the normal equation for any B_j term can be done in 7 steps:

   1. drop the exponent 2 and set function in phase B equal to 0
   2. multiply the function by Z_j
   3. distribute the Z_j term
   4. distribute the summation operator
   5. apply rules of summation for constants
   6. solve in terms of the criterion variable, Z_Y
   7. apply rules for standard scores and simplify.

Applying these steps in turn produces:

1. $\sum(Z_Y - A - B_1Z_1 - B_2Z_2 - B_3Z_3 -\ldots- B_jZ_j -\ldots- B_pZ_p) = 0$

2. $\sum(Z_Y - A - B_1Z_1 - B_2Z_2 - B_3Z_3 -\ldots- B_jZ_j -\ldots- B_pZ_p)Z_j = 0$

3. $\sum(Z_YZ_j - AZ_j - B_1Z_1Z_j - B_2Z_2Z_j - B_3Z_3Z_j -\ldots- B_jZ_j^2 -\ldots- B_pZ_jZ_p) = 0$

4. $\sum Z_YZ_j - \sum AZ_j - \sum B_1Z_1Z_j - \sum B_2Z_2Z_j - \sum B_3Z_3Z_j -\ldots- \sum B_jZ_j^2 -\ldots- \sum B_pZ_jZ_p = 0$

$$5.\ \sum z_y z_j - A\sum z_j + B_1\sum z_1 z_j + B_2\sum z_2 z_j + B_3\sum z_3 z_j + \ldots + B_j\sum z_j^2 + \ldots + B_p\sum z_j z_p = 0$$

$$6.\ \sum z_y z_j = A\sum z_j + B_1\sum z_1 z_j + B_2\sum z_2 z_j + B_3\sum z_3 z_j + \ldots + B_j\sum z_j^2 + \ldots + B_p\sum z_j z_p$$

$$7.\ (n-1)r_{yj} = 0 + B_1(n-1)r_{1j} + B_2(n-1)r_{2j} + B_3(n-1)r_{3j} + \ldots + B_j(n-1) + B_p(n-1)r_{jp}$$

Dividing through by $(n-1)$,

$$r_{yj} = B_1 r_{1j} + B_2 r_{2j} + B_3 r_{3j} + \ldots + B_j + \ldots + B_p r_{jp}$$

Alternate Procedure

The above normal equation (r_{yj}) is a general result for any B_j term. Now a much simpler method for finding a full set of normal equations will be discussed.

Recall that the correlation of a variable with itself is equal to 1. That is, $r_{11} = r_{22} = r_{33} = \ldots = r_{pp} = 1$. From this fact and the general normal equation, one can write out the entire set of normal equations for any number of predictors. If r_{yj} holds for any B_j term, then it holds for $j=1, j=2, j=3, \ldots, j=p$. For example, assume $p=2$ predictors. We know that the set of normal equations will consist of $(p)(p) = 2^2$ terms. Thus, first first write out r_{yj} twice as follows:

$$r_{yj} = B_1 r_{1j} + B_2 r_{2j}$$

$$r_{yj} = B_1 r_{1j} + B_2 r_{2j}$$

Now substitute the appropriate j value: j=1 throughout for line 1, and j=2 throughout for line 2:

$$r_{y1} = B_1 r_{11} + B_2 r_{21} \qquad r_{y1} = B_1 + B_2 r_{21}$$

or

$$r_{y2} = B_1 r_{12} + B_2 r_{22} \qquad r_{y2} = B_1 r_{12} + B_2$$

or

$$r_{y1} = B_1 + B_2 r_{12}$$

$$r_{y2} = B_1 r_{12} + B_2$$

The third set shows the subscripts in reverse order for the B term correlatio

$r_{21} = r_{12}$. Although not necessary, it may be easier to do this in order to conform to this convention as we have done in the derivations.

In summary, finding normal equations for any number of predictors involves the eated application of several steps for each of the B_j terms.

Example for Five Predictors

To exemplify the procedure for p predictors, we will work through the solution of normal equations for five predictors. We will first do it by the long method, and then show a solution by the shorter method.

A. The regression model is:

$$Z_{\hat{Y}} = A + B_1 Z_1 \quad B_2 Z_2 + B_3 Z_3 + B_4 Z_4 + B_5 Z_5$$

B. The function to be minimized is:

$$\sum (Z_Y - A - B_1 Z_1 - B_2 Z_2 - B_3 Z_3 - B_4 Z_4 - B_5 Z_5)^2$$

C. The normal equation for A =0. To derive the normal equations for B_1 through B_5 apply the 7 steps listed on page 32 used for finding the normal equation for any B_j term. The amount of algebra involved in doing this requires an efficient proced One method that may be found useful is as follows:

1. write step 1 as $\sum (Z_Y - Z_{\hat{Y}}) = 0$

2. write the second step for each of the B_j terms as:

$$\sum(Z_Y - Z_{\hat{Y}})Z_1 = 0$$
$$\sum(Z_Y - Z_{\hat{Y}})Z_2 = 0$$
$$\sum(Z_Y - Z_{\hat{Y}})Z_3 = 0$$
$$\sum(Z_Y - Z_{\hat{Y}})Z_4 = 0$$
$$\sum(Z_Y - Z_{\hat{Y}})Z_5 = 0$$

3. write step 3 for each of the terms as:

$$\sum(Z_Y Z_1 - Z_1 Z_{\hat{Y}}) = 0$$
$$\sum(Z_Y Z_2 - Z_2 Z_{\hat{Y}}) = 0$$
$$\sum(Z_Y Z_3 - Z_3 Z_{\hat{Y}}) = 0$$
$$\sum(Z_Y Z_4 - Z_4 Z_{\hat{Y}}) = 0$$
$$\sum(Z_Y Z_5 - Z_5 Z_{\hat{Y}}) = 0$$

4. steps 4, 5, and 6 can be combined to yield:

$$\sum Z_Y Z_1 = \sum(B_1 Z_1 + B_2 Z_2 + B_3 Z_3 + B_4 Z_4 + B_5 Z_5)Z_1$$
$$\sum Z_Y Z_2 = \sum(B_1 Z_1 + B_2 Z_2 + B_3 Z_3 + B_4 Z_4 + B_5 Z_5)Z_2$$
$$\sum Z_Y Z_3 = \sum(B_1 Z_1 + B_2 Z_2 + B_3 Z_3 + B_4 Z_4 + B_5 Z_5)Z_5$$
$$\sum Z_Y Z_4 = \sum(B_1 Z_1 + B_2 Z_2 + B_3 Z_3 + B_4 Z_4 + B_5 Z_5)Z_4$$
$$\sum Z_Y Z_5 = \sum(B_1 Z_1 + B_2 Z_2 + B_3 Z_3 + B_4 Z_4 + B_5 Z_5)Z_5$$

5. distribute the Z_j term and the summation operator :

$$\sum Z_Y Z_1 = B_1\sum Z_1^2 + B_2\sum Z_1 Z_2 + B_3\sum Z_1 Z_3 + B_4\sum Z_1 Z_4 + B_5\sum Z_1 Z_5$$
$$\sum Z_Y Z_2 = B_1\sum Z_1 Z_2 + B_2\sum Z^2 + B_3\sum Z_2 Z_3 + B_4\sum Z_2 Z_4 + B_5\sum Z_2 Z_5$$
$$\sum Z_Y Z_3 = B_1\sum Z_1 Z_3 + B_2\sum Z_2 Z_3 + B_3\sum Z_3^2 + B_4\sum Z_3 Z_4 + B_3\sum Z_3 Z_5$$
$$\sum Z_Y Z_4 = B_1\sum Z_1 Z_4 + B_2\sum Z_2 Z_4 + B_3\sum Z_3 Z_4 + B_4\sum Z_4^2 + B_5\sum Z_4 Z_5$$
$$\sum Z_Y Z_5 = B_1\sum Z_1 Z_5 + B_2\sum Z_2 Z_5 + B_3\sum Z_3 Z_5 + B_4\sum Z_4 Z_5 + B_5\sum Z_5^2$$

Applying rules for standard scores:

$$(n-1)r_{y1} = B_1(n-1) + B_2(n-1)r_{12} + B_3(n-1)r_{13} + B_4(n-1)r_{14} + B_5(n-1)r_{15}$$

$$(n-1)r_{y2} = B_1(n-1)r_{12} + B_2(n-1) + B_3(n-1)r_{23} + B_4(n-1)r_{24} + B_5(n-1)r_{25}$$

$$(n-1)r_{y3} = B_1(n-1)r_{13} + B_2(n-1)r_{23} + B_3(n-1) + B_4(n-1)r_{34} + B_5(n-1)r_{35}$$

$$(n-1)r_{y4} = B_1(n-1)r_{14} + B_2(n-1)r_{24} + B_3(n-1)r_{34} + B_4(n-1) + B_5(n-1)r_{45}$$

$$(n-1)r_{y5} = B_1(n-1)r_{15} + B_2(n-1)r_{25} + B_3(n-1)r_{35} + B_4(n-1)r_{45} + B_5(n-1)$$

Dividing through by (n-1) produces the simplified set of normal equations:

$$r_{y1} = B_1 + B_2r_{12} + B_3r_{13} + B_4r_{14} + B_5r_{15}$$

$$r_{y2} = B_1r_{12} + B_2 + B_3r_{23} + B_4r_{24} + B_5r_{25}$$

$$r_{y3} = B_1r_{13} + B_2r_{23} + B_3 + B_4r_{34} + B_5r_{35}$$

$$r_{y4} = B_1r_{14} + B_2r_{24} + B_3r_{34} + B_4 + B_5r_{45}$$

$$r_{y5} = B_1r_{15} + B_2r_{25} + B_3r_{35} + B_4r_{45} + B_5$$

The short method for 5 predictors begins by writing the 5^2 terms for the general r_{yj} normal equation. See the next page.

$$r_{yj} = B_1 r_{1j} + B_2 r_{2j} + B_3 r_{3j} + B_4 r_{4j} + B_5 r_{5j}$$

$$r_{yj} = B_1 r_{1j} + B_2 r_{2j} + B_3 r_{3j} + B_4 r_{4j} + B_5 r_{5j}$$

$$r_{yj} = B_1 r_{1j} + B_2 r_{2j} + B_3 r_{3j} + B_4 r_{4j} + B_5 r_{5j}$$

$$r_{yj} = B_1 r_{1j} + B_2 r_{2j} + B_3 r_{3j} + B_4 r_{4j} + B_5 r_{5j}$$

$$r_{yj} = B_1 r_{1j} + B_2 r_{2j} + B_3 r_{3j} + B_4 r_{4j} + B_5 r_{5j}$$

Substitute the appropriate j term (j=1 for line 1, j=2 for line 2, etc.) and set $r_{11}=1$, $r_{22}=1$, etc.

$$r_{y1} = B_1 + B_2 r_{21} + B_3 r_{31} + B_4 r_{41} + B_5 r_{51}$$

$$r_{y2} = B_1 r_{12} + B_2 + B_3 r_{32} + B_4 r_{42} + B_5 r_{52}$$

$$r_{y3} = B_1 r_{13} + B_2 r_{23} + B_3 + B_4 r_{43} + B_5 r_{53}$$

$$r_{y4} = B_1 r_{14} + B_2 r_{24} + B_3 r_{34} + B_4 + B_5 r_{54}$$

$$r_{y5} = B_1 r_{15} + B_2 r_{25} + B_3 r_{35} + B_4 r_{45} + B_5$$

Now, if one desire, the subscripts of the predictor correlations can be reversed such that the first subscript is less than the second. The result is the same set of normal equations derived under the long method.

NOTE

Three related papers that may be of interest to the reader are in preparation for publication in the ERIC system. Tentative titles and expected order of appearance are as follows:

1. A derivation of the sample multiple correlation formula for raw scores.

2. A derivation of the unbiased sample standard error of estimate.

3. A matrix algebra technique for finding beta weights from a a correlation matrix.

References

O'Brien, Francis J., Jr. A proof that t^2 and F are identical: the general case. ERIC Clearinghouse on Science, Mathematics and Environmental Education, Ohio State University, 1982a. ED 215894.

_____. Proof that the sample bivariate correlation coefficient has limits $\pm$ 1. ERIC Clearinghouse on Science, Mathematics, and Environmental Education, Ohio State University, 1982b. ED 216874.